Data Mining Classification Models for Industrial Planning

Ricardo Bragança[1], Filipe Portela[1(✉)], A. Vale[2], Tiago Guimarães[1], and Manuel Santos[1]

[1] Algoritmi Research Centre, University of Minho, Guimarães, Portugal
a5l055@alunos.uminho.pt, {cfp,mfs}@dsi.uminho.pt
[2] Value Added Partners, Porto, Portugal

Abstract. The data mining models are an excellent tool to help companies that live from the sale of items they produce. With these models combined with Lean Production, it becomes easier to remove waste and optimize industrial production. This project is based on the phases of the methodology CRISP-DM. Several methods were applied to this data namely, average, mean and standard deviation, quartiles and Sturges rule. Classification Techniques were used in order to understand which model has the best probability of hitting the correct result. After performing the tests, model M1 was the one with the best chance to accomplish a great level of classification having 99.52% of accuracy.

Keywords: Data mining · Classification · CRISP-DM · DSR · Lean · WEKA

1 Introduction

Companies in industry are increasingly feeling the need to find a way to optimize their production to meet the adversities of the economic world. One of the best ways to do this is using the data mining models that allow, based on past sales, get an estimate of how much will sell the right time, production efficiently and reducing the waste of raw material and labor work. The rating models are a great tool to help businesses achieved success. Having said that, test yourself several models in terms of classification hoping to find models with higher accuracy than 90%.

This paper initially made an overview of the concepts inherent in the project including data mining, Lean Production and Decision Support ending with what exists today. Subsequently it is presented methodologies, CRISP-DM, DSR and at the level of the tools was used WEKA. Then the paper presents the results following the phases of the CRISP-DM. Finally, it created a little discussion about the project and concluding with the presentation of closing arguments.

M. Singh et al. (Eds.): ICACDS 2016, CCIS 721, pp. 585–594, 2017.
DOI: 10.1007/978-981-10-5427-3_60

2 Background

2.1 Data Mining

In the world there is a vast amount of data stored, but these are only examined in a very superficial way, which leads to having a wealth of data and great poverty of knowledge [1]. In the last decades, data mining has been widely recognized as an powerful and versatile tool of data analysis and has made a significant contribution in the areas of information technology clinical medicine, sociology, physics, in the areas of management, economics and finance [2]. Data mining (DM) is the task that seeks to discover patterns in data sets by using methods of artificial intelligence, machine learning, statistics and database systems [3]. Wu [4] suggests a more complete definition for DM, he believes it is the integration of various subjects: (i) databases, (ii) databases technologies, (iii) statistic, (iv) machine learning, (v) math, (vi) neural networks and others. The DM was designed to use two techniques, predictive and descriptive. The predictive also known as supervised learning is characterized by obtaining examples and input data, predicting future values. In turn, descriptive also known as unsupervised learning is characterized by learning through data grouping with identical characteristics [5].

2.2 Lean Production (LP)

LP originates from the Toyota Company, started in the end of World War II with the implementation of the Toyota Production System (TPS) and aimed to increase productivity and reduce the cost of car production by eliminating any waste. The fact that this system allowed for Toyota leadership in the automotive market has led researchers at the Massachusetts Institute of Technology (MIT) dubs it Lean Production [6]. The LP is an organizational model that brings numerous benefits to the organization that implements it by reducing costs and eliminating waste [7]. There is a great difficulty to find the best setting for LP as the existing definitions and their focus differs from author to author, as well as the practices involved [8]. The Lean term comprises the establishment of a culture of continual improvement and organizational learning [9]. At the operating level, it reduces unnecessary production time, materials and manufacturing efforts, thus reducing the cost by reducing waste [10]. There are obstacles that must be taken into account when seeking the implementation of this methodology, namely the human factor, often brings the rejection of change by employees and that can cast doubt to the implementation of Lean [11]. The implementation of Lean is a lengthy process, thus requiring a well-defined strategy and must involve top management in the process so that there is real knowledge of the organization [6].

2.3 Decision Support

The Decision Support refers to applications involving broad analyses and exploration of current and historical data in organizations supporting the high-level

Decision-making. The decision Support requires the integration of two types of management:

- **Data management,** which includes the organization's databases that are managed by database management systems;
- **Knowledge management,** which handles tasks related to reasoning.

Due to various limitations existing in the processing of human Knowledge it is necessary to use other support. In the Simon's decision-making process there are three main phases [12]:

- **Intelligence,** at this stage the decision maker observes reality, identifies and defines the problems. The main requirement of decision support is the ability to observe internal and external sources of information in search of opportunities and problems interpreting being it discovered.
- **Design,** here is built the representative model of the system through assumptions that simplify the reality and describe the relationship between variables, then the model is validated. In the design phase, the support to decision involves generating alternatives, discussion of selection criteria, the importance of the chosen criteria, and forecasting consequences.
- **Choice,** at this stage it is made the choice of model of the proposed solution, following its assessment. Different scenarios are tested in order to select the option that reinforces the final decision in decision support.

2.4 Related Work

In the production context, new sensor technologies and the increased application of simulation and monitoring systems led to an enormous increase of manufacturing data. Additionally, a new approach for the assessment of manufacturing quality based on process signals from the machine tool is proposed, which provides current tool state and information for every manufacturing process. In order to reuse and evaluate this data for knowledge-based process planning, an approach to manufacturing data collection and evaluation using data mining methods was developed. An analysis and classification of manufacturing data has been carried out to identify input data for knowledge-based process planning [13]. Ming-Te et al. [14], suggest the use of DM, notably Genetic Programming, Artificial Neural Networks and Logistic Regression to improve the LP configuration for improved performance.

Groger et al. [15] presented a paper describing the optimization of manufacturing as a targeted approach to the DM. Based on a huge amount of data, previously defines the cases of DM use that will be applied to identify hidden data patterns to optimize the entire manufacturing process. By using Dashboards and Key Performance Indicators (KPI) it is possible to analyze in the depth the reasons for deviations and presented indication in order to improve processes.

Unver [16] indicates that there is an opportunity expressed in manufacturing industries for a concept denominated Intelligent manufacturing. This concept is used all the advantages of Lean: cost control; quality improvement and product and waste

disposal, combined with business intelligence more precisely the dashboards presenting the processed data and KPI, which allow managers to determine the most appropriate options for applying Lean. Vazan et al. [17] indicates that there is an opportunity expressed in manufacturing industries for a concept denominated Intelligent manufacturing. This concept is used all the advantages of Lean: cost control; quality improvement and product and waste disposal, combined with business intelligence more precisely the dashboards presenting the processed data and KPI, which allow managers to determine the most appropriate options for applying Lean.

3 Methods and Tools

Throughout the project, we used two methodological approaches, Design Science Research (DSR) and Cross Industry Standard Process for Data Mining (CRISP-DM).

As the first methodology is in the field of Information Systems (IS), it was used for conducting the research process. As the second one is a project linked to data mining it will be used in carrying out the project. In the last years the DSR regained importance in the paradigm of research specifically in the field of IS [18]. In the IS area, DSR is used in the construction of artifacts in a countless number of areas from systems for decision support, Modeling tools, assessment methods and among others [19]. DSR is a set of techniques and synthetic and analytical perspectives to conduct research processes. The main objective of this approach is the creation of knowledge through the design of new and innovative artifacts and analysis of their use and performance along with reflection and abstraction [20]. DSR consists of six phases: (i) Identification of the problem and motivation, which defines the specific research problem and where the value of the solution is justified; (ii) Definition of the solution objectives and identification of goals for the problem definition and knowledge of what is possible; (iii) Development and Design of the solution, at this stage artifacts for the solution are created; (iv) Demonstration, is intended at this stage to demonstrate the efficacy of articles developed in the previous stage to solve the problem; (v) Evaluation, at this stage is intended to observe and measure how the artifact can actually support the solution of the problem; (vi) Communication, at this stage communication is made to the researchers and/or professionals interested, what is the found problem, its importance, which artifacts have been developed, as well as its usefulness and the fact of whether it is a novelty or not [21].

One of the main methods of data mining is the CRISP-DM and it was developed by a consortium comprised of NCR Systems Engineering Copenhagen, DaimlerChrysler AG, SPSS Inc. and OHRA Verzekering en Bank Groep B.V. [22]. The main objective of this methodology is to define a complete data mining process from Understanding the data for the implementation of developed models through the monitoring of improvements The main objective of this methodology consists in defining a complete data mining process from understanding the data to the implementation of developed models through the monitoring of improvements [23].

This approach consists of six main phases: (1) Business Understanding, (2) Data Understanding, (3) Data Preparation, (4) Modeling, (5) Evaluation and (6) Implementation [24]. The WEKA was the tool chosen to implement this project because despite being widely used in data mining projects is also open source.

Taking into account the circumstances of the project, it can be said that an investigation is centered on a problem, which leads to starting the investigation on the problem identification and motivation stage at the DSR level and finish in the communication phase. The problem that gave rise to this research is the existence, in the organization to be held, at the level of intervention, of a misfit production in the face of orders received as excess produce, which entails a series of costs. Through the development of the article, we intend to adjust the production to the reality of the company as well as the reduction/elimination of existing waste. Therefore, it is essential to use the CRISP-DM methodology in order to understand how the business works, understand the data, and prepare the data, eliminating what does not make sense, obtaining the models that were applied to the data and evaluation of models. The evaluation was performed using WEKA tool that helped get the results of each model.

4 Industrial Planning

As previously stated the entire project was developed following CRISP-DM methodology. The next section describes the stages of this methodology.

4.1 Business Understanding

This project is part of the need to eliminate all types of waste inherent in the process of a producer organization of textile labels and their role is to design, produce, manipulate and encode all types of labels for de leading companies of confection and sale of textile products and supplements worldwide. For this purpose, we developed data mining models that will be presented in subsequent points.

4.2 Data Understanding

The data used to carry out this project concern the quantity sold of a product from 09-11-2012 to 19-02-2016 and contain 4299 records. In Table 1 is found a description of the attributes that initially existed before proceeding to any kind of change. After performing a data analysis, it was denoted the existence of periods with no data because there was no data from the beginning to the end of the calendar year analysis.

Table 1. Description of the initial attributes

Attribute	Description	Example	Maximum	Minimum	Average
Article	Article nomenclature	REFXPTA	n.a.	n.a.	n.a.
Data	Day, month and year in which particular item has been sold	2014-01-06	19-02-2016	09-11-2012	n.a.
Quantity	Quantity sold by a particular article	47.350.692	199.000	−31.800	7.273.498

4.3 Data Preparation

In order to obtain results as reliable as possible the interval was shortened from 01-01-2014 to 30-12-2015, thereby obtaining 52 completed weeks in each of the years. Given the introduction of the number of weeks, in addition to changing the attributes, they were also performed calculations in the attribute of quantity sold, particular sum and medium, thus obtaining the classes to which the quantities belong. In Table 2 is a description of the attributes that will be used later in the models.

Table 2. Description of the attributes that have changed

Attribute	Description	Example
Article	Article nomenclature	REFXPTA
Year	Year in which a particular item has been sold	2014
Week	Number of the week a particular item has been sold	1
Quantity	Sold quantity of a certain item	196.745
Class	Class is inserted where the quantities	]0;196.745]

After completing the grouping, it was necessary to resort for methods able to identify the quantity sold interval, to perform predictions using the classification approach.

Taking into account the absence of a method able to determine these intervals, it was used: Average, quartiles, medium and standard deviation and Sturges rule.

The average was calculated using the mathematical expression $\overline{\chi} = \frac{1}{n}\sum_{i=1}^{n} xi$ where n is the number of weeks and xi is the quantity sold of given article. The intervals are $[min, \bar{x}]$ and $]\bar{x}, max]$. Another important method is the quartiles, a non- central measure of tendency, that in face of a sample, is intended to determine a set of four distinct classes. In the first instance, it was necessary to sort the data set in ascending order and then identify the maximum and minimum value. To determine the first class is necessary to identify the minimum quantity sold of article sample REFXPTA and the value of the sold amount corresponding to the first quartile to determine this amount was necessary to use the following calculation $xn = \frac{(25\% * n)}{100\%}$. The first class has defined as a range between two values, $\left[min, xn = \frac{25\% * n}{100\%}\right]$, the range determines the following class $\left]xn = \frac{25\% * n}{100\%}, xn = \frac{50\% * n}{100\%}\right]$, the third class is defined by the range $\left]xn = \frac{50\% * n}{100\%}, xn = \frac{75\% * n}{100\%}\right]$ and finally the fourth grade $\left]xn = \frac{75\% * n}{100\%}, max\right]$. The mean and standard deviation methods require the same calculation made in the first describe method, but it needs the value of the standard deviation. Once identified, its value was determined by:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(xi - \bar{x})^2}{n-1}}$$

The variable xi is the quantity sold per week and n the number of weeks. From the determined values, the classes were defined. One of the intervals can be set by $]\bar{x}, \bar{x} + s]$,

but it is important to note that the upper and lower values cannot exceed the *max* and *min* respectively. Finally, we determine the classes using the Sturges rule. The number of classes is determined from the expression $k = 1 + 3,322 * \log_{10} 208$ and the amplitude of each class was determined by reference to $amplitude = (max - \min)/k$. The first class corresponds to the range that is determined by the *min* value calculation and application $h1 = \min + amplitude$, The first range is obtained from $k1 = [min, h1]$. The second range will be obtained from $k2 =]h1, h1 + amplitude]$; the remaining intervals are calculated according to this process, sequentially to determine the set of class. In Table 3 is an example of dataset, which corresponds to the average data amount sold, by applying the Sturges rule.

Table 3. Example dataset

Class number	Class	Quantity
1	[0;59535875.13]	19
2	]59535875.13;119071750.3]	26
3	]119071750.3;178607625.4]	31
4	]178607625.4;238143500.5]	10
5	]238143500.5;297679375.625]	8
6	]297679375.625;357215250.75]	2
7	]357215250.75;416751125.875]	4
8	]416751125.875;476287000]	4

4.4 Modeling

While carrying out this project and having in consideration that it intends to build predictive models in order to foresee what it will produce, supervised learning techniques will be used, more precisely in this case the classification. In the classification, it was used as a target to classes obtained by applying the methods referenced from above. The techniques used in the models were classification Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB). The techniques used to the classification level were J48, Naive Bayes, MultyLayerPerceptron e LisbSVM (Table 4). The sampling method 10-folds Cross Validation was implemented to test mechanism for the models. At the level of NB, a default application was utilized.

Table 4. Description of models

Model	Description
M1	Average application to the average quantities sold
M2	Application quartile average of quantities sold
M3	Application average and standard deviation of the average quantities sold
M4	Application Sturges rule to the average quantities sold
M5	Average application to the sum of quantities sold
M6	Application of quartiles to the sum of quantities sold
M7	Application the mean and standard deviation to the sum of quantities
M8	Application Sturges rule to the sum of quantities sold

Regarding the SVM, a radial basis function at the level of the kernel was used. The DT's possessed a confidence factor of 0.25 and a number of folds equals to three.

4.5 Evaluation

In order to determine which is the best model for use in industrial planning, it was carried out their evaluation for the classification (Table 5). According to the accuracy, which is the ability of a classifier correctly predict the class of new data and the accuracy average, as its name indicates is the average of the classifiers applied to each model. The best models are those in which the percentage of accuracy is closer to 100%. Following this, the models M1 and M5 with a percentage of accuracy of 99.52% and 86.54% respectively, which means that these designs are more likely to achieve. At the other end of the standings, have the M2 and M4 models with 72.84% e 74.04% respectively. Note that the four algorithms tested in each model, the more times that obtained better result was J48 decision tree followed by Naive Bayes.

Table 5. Classification results

Model	Best algorithm	Accuracy	Accuracy average
M1	MultyLayer Perceptron; LisbSVM	100.00%	99.52%
M2	J48	97.12%	72.84%
M3	J48	94.23%	80.05%
M4	NaiveBayes	91.35%	74.04%
M5	J48	99.04%	86.54%
M6	J48	97.12%	75.96%
M7	J48; NaiveBayes	96.15%	82.21%
M8	NaiveBayes	96.15%	77.89%

5 Discussion

After completing the analysis of the obtained models, it could be perceived that on average the specimens had higher percentage of accuracy, in other words, those that are more likely to hit, are the models that have fewer classes as target. In this case, both the average the sum of the quantity sold as the average number of average sold. However, only the second can achieve accuracy higher than the acceptable values (95%). On the other end, the classification model that contained the average quantity sold per week, which was applied, to Sturges rule was what got worst rating with about 91% accuracy.

On the subject of acuity, the best designs are repeated with the model M1 on the lead followed by M5. Regarding the last classified, there were scrambled, being in last place also the average quantities sold, but this time it was applying the standard deviation. On average, the algorithm that obtained better result more often at the level of acuity was the one belonging to decision trees J48.

6 Conclusion and Future Work

To conclude, it is important to understanding how the data mining models are significant in industrial planning. Together with methodologies such as Lean Production, is a fundamental tool to combat waste and help the optimization of industrial production. Taking into account the data obtained and its modifications it can be shown that, at the level of classification, the best model is the M rating because it is what had the best percentage of accuracy meaning that it is more likely to hint.

When it comes to forecasting quantities sold, it is no enough just to look at the quantities sold in previous years it is also necessary to take into account the weather to realize the impact that it may have had on production.

In the future work, it is important to carry out the last phase of CRISP-DM, which is the phase of implementation, taking these models into practice, helping this way all organization to improve their production process.

Today there is a huge amount of production data (Big Data), and each time more, people want to get the necessary information in real time, which requires us to learn and adapt our technologies and techniques. Given this, it was interesting to explore these increasingly dominant options in the world of technology.

Acknowledgements. This work has been supported by Compete: POCI-01-0145-FEDER-007043 and FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

References

1. Alsultanny, Y.: Labor market forecasting by using data mining. Procedia Comput. Sci. **18**, 1700–1709 (2013)
2. Xu, W., Zheng, T., Li, Z.: A neural network based forecasting method for the unemployment rate: prediction using the search engine query data. Presented at the 2011 Eighth IEEE International Conference on e-Business Engineering (2011)
3. Ramos, S., Duarte, J., Duarte, F.J., Vale, Z.: A data-mining-based methodology to support MV electricity customers characterization. Procedia Energy Build. **91**, 16–25 (2015)
4. Yan, W.: Application research of data mining technology about teaching quality assessment in colleges and universities. Procedia Eng. **15**, 4241–4245 (2011)
5. Ren, X., Yan, D., Hong, T.: Data mining of space heating system performance in affordable housing. Procedia Build. Environ. **89**, 1–13 (2015)
6. Maia, L., Alves, A., Leão, C.: Metodologias para Implementar Lean Production: Uma Revisão Crítica de Literatura. Presented at the 6º Congresso Luso-Moçambicano de Engenharia (CLME2011) A Engenharia no combate à pobreza, pelo desenvolvimento e competitividade (2011)
7. Maia, L., Alves, A., Leão, C.: Definition of a protocol for implementing lean production methodology in textile and clothing case studies. Presented at the ASME 2013 International Mechanical Engineering Congress and Exposition, San Diego, California, USA (2013)
8. Hasle, P., Bojesen, A., Jensen, P., Bramming, P.: Lean and working environment: a review of the literature. Int. J. Oper. Prod. Manag. **32**(7), 829–849 (2012)

9. Yamamoto, Y., Bellgran, M.: Fundamental mindset that drives improvements towards lean production. Assem. Autom. **30**(2), 124–130 (2010)
10. Hassan, K., Kajiwara, H.: Application of pull concept-based lean production system in the ship building industry. J. Ship Prod. Des. **29**(3), 105–116 (2013)
11. Balashova, E., Gromova, E.: Prospects and specifics of resource management in enterprises operating in different sectors of the Russian economy. Econ. Manag. Enterp. **216**(2), 102–108 (2015)
12. Turban, E., Sharda, R., Delen, D.: Decision Support and Business Intelligence Systems, 9th edn. Prentice Hall, Upper Saddle River (2011)
13. Denkena, B., Schmidt, J., Kruger, M.: Data mining approach for knowledge-based process planning. Presented at the 2nd International Conference on System-Integrated Intelligence: Challenges for Product and Production Engineering (2014)
14. Ming-Te, L., Kuo-Chung, M., Pan, W.-T.: Using data mining technique to perform the performance assessment of lean service. Neural Comput. Appl. **22**(7), 1433–1445 (2013)
15. Groger, C., Nidermann, F., Mitschang, B.: Data mining-driven manufacturing process optimization. Presented at the World Congress on Engineering 2012, London, UK (2012)
16. Unver, H.: An ISA-95-base manufacturing intelligence system in support of lean initiatives. Int. J. Adv. Manuf. Technol. **65**(5), 853–866 (2013)
17. Vazan, P., Tanuska, P., Kebisek, M.: The data mining usage in production system management. Int. J. Mech. Aerosp. Ind. Mechatron. Manuf. Eng. **5**(5), 922–926 (2011)
18. Myers, M., Venable, J.: A set of ethical principles for design science research in information systems. Procedia Inf. Manag. **51**, 801–809 (2014)
19. Gregor, S., Hevner, A.: Positioning and presenting design science research for maximum impact. MIS Q. **37**(2), 337–355 (2013)
20. Vaishnavi, V., Kuechler, B.: Design science research in information systems (2013). http://desrist.org/desrist/content/design-science-research-in-information-systems.pdf
21. Hain, S., Andrea, B.: Towards a maturity model for e-collaboration - a design science research approach. Presented at the 44th Hawaii International Conference on System Sciences (2011)
22. Erohin, O., Kuhlang, P., Schallow, J., Deuse, J.: Intelligent utilisation of digital databases for assembly time determination in early phases of product emergence. Procedia CIRP **3**, 424–429 (2012)
23. Hoe, A., et al.: Analyzing students records to identify patterns of students performance. Presented at the 2013 International Conference on Research and Innovation in Information Systems (ICRIIS) (2013)
24. Wallis, R., Erohin, O., Klinkenberg, R., Deuse, J., Stromberger, F.: Data mining - supported generation of assembly process plans. Procedia CIRP **23**, 178–183 (2014)